**MODULE 4**

**Key Statistical Concepts** for Data Science

Statistics provides the mathematical foundation for analyzing, interpreting, and making decisions from data.Below are the essential concepts every data scientist should know:

**1. Types of Data**

* **Qualitative (Categorical)** – Represents categories or labels.  
  Example: Gender, Country, Product type.
* **Quantitative (Numerical)** – Represents measurable quantities.
  + **Discrete**: Countable values (e.g., number of customers).
  + **Continuous**: Measurable values within a range (e.g., height, weight).

**2. Descriptive Statistics**

Used to **summarize and describe** data.

* **Measures of Central Tendency**
  + **Mean**: Average of values.
  + **Median**: Middle value when data is sorted.
  + **Mode**: Most frequent value.
* **Measures of Dispersion (Spread)**
  + **Range**: Difference between max and min.
  + **Variance**: Average squared deviation from the mean.
  + **Standard Deviation (SD)**: Square root of variance, shows spread in same units as data.
  + **Interquartile Range (IQR)**: Spread of middle 50% of data.

**3. Probability**

Probability measures the **likelihood** of an event occurring.

* **0** means impossible, **1** means certain.
* **Basic rules**:
  + P(A∪B)=P(A)+P(B)−P(A∩B)P(A \cup B) = P(A) + P(B) - P(A \cap B)P(A∪B)=P(A)+P(B)−P(A∩B)
  + P(A∣B)=P(A∩B)P(B)P(A | B) = \frac{P(A \cap B)}{P(B)}P(A∣B)=P(B)P(A∩B)​ (Conditional probability)

**4. Distributions**

Statistical distributions describe how values are spread.

* **Normal Distribution** (bell curve) – Common in natural data.
* **Binomial Distribution** – Success/failure experiments.
* **Poisson Distribution** – Count of events in a fixed time/space.
* **Uniform Distribution** – All outcomes equally likely.

**5. Inferential Statistics**

Drawing **conclusions about a population** from a sample.

* **Hypothesis Testing**
  + **Null Hypothesis (H0H\_0H0​)** – No effect or difference.
  + **Alternative Hypothesis (HaH\_aHa​)** – Effect exists.
  + **p-value** – Probability of observing data assuming H0H\_0H0​ is true.  
    If p<0.05p < 0.05p<0.05, reject H0H\_0H0​.
  + **Confidence Interval (CI)** – Range in which true value lies with a certain probability (e.g., 95% CI).
* **Common Tests**
  + **t-test** – Compare means between groups.
  + **Chi-square test** – Compare categorical distributions.
  + **ANOVA** – Compare means of 3+ groups.
  + **Correlation (Pearson, Spearman)** – Measure strength of relationship.

**6. Correlation vs. Causation**

* **Correlation**: Two variables move together.
* **Causation**: One variable directly affects the other.
* Data science often requires experiments or domain knowledge to prove causation.

**7. Sampling Methods**

* **Random Sampling** – Every member has equal chance.
* **Stratified Sampling** – Sample proportionally from subgroups.
* **Systematic Sampling** – Select every k-th member.
* **Cluster Sampling** – Select entire groups (clusters) randomly.

**8. Outliers and Skewness**

* **Outliers**: Data points far from the rest.
* **Skewness**: Measure of asymmetry in distribution (positive or negative).

**Summary:**

Understanding these statistical concepts helps data scientists summarize, analyze, and draw meaningful conclusions from data while avoiding wrong interpretations.